REMARKS

Claims 1-36 were pending in this application. In an Office Action dated January 17, 2007, claims 1-36 were rejected. Applicants thank the Examiner for examination of the claims pending in this application and address the Examiner's comments below.

Applicants are canceling claims 2, 7, 15, and 26 and amending claims 1, 6, 10, 12, 13, 14, 24, 28 and 36 in this Amendment and Response. These changes are not believed to introduce new matter, and their entry is respectfully requested. In making these amendments, Applicants do not concede that the subject matter of such claims was in fact disclosed or taught by the cited prior art. Rather, Applicants reserve the right to pursue such protection at a later point in time and merely seek to pursue protection for the subject matter presented in this submission.

In view of the Amendments herein and the Remarks that follow, Applicants respectfully request that the Examiner reconsider all outstanding objections and rejections, and withdraw them.

Response to Objections to the Specification

In paragraph 2 of the Office Action, the Examiner objects to the disclosure because of an alleged informality. The Examiner asserts that "$L(H_c)$" should be changed to "$L(H_i)$" on page 12 line 22. This objection is traversed.

In the Applicants' specification, lines 18-23 discuss the calculation of the collocation hypothesis, $L(H_c)$ which is defined in the following equation (3). The use of $L(H_c)$ on page 12, line 22 is intentional and correct.

Response to Rejection Under 35 USC § 112, Paragraph 2

In the 3rd, 4th and 5th paragraphs of the Office Action, the Examiner rejects claims 13 and

14 under 35 USC 112, Paragraph 2 as allegedly failing to point out and distinctly claim the

subject matter which Applicants regard as invention. Specifically, the Examiner asserts that

claims 13 and 14 lack sufficient antecedent basis.

Independent claim 13 has been amended to recite "the limit for the iteration" for which

there sufficient antecedent basis in the preceding claim elements. Claim 14 has been amended to

recite "the stored upper limit" for which there is sufficient antecedent basis in preceding claim

elements. Thus, Applicants submit that claims 13 and 14 point out and distinctly claim the

subject matter which Applicants regard as invention.

Response to Rejection Under 35 USC 103(a)

In the 6th, 7th and 8th paragraphs of the Office Action, the Examiner rejects claims 1, 3, 6,

8, 11-13, 20-24 and 31-36 under 35 USC 103(a) as allegedly being unpatentable over Su et al.

(In Proceedings of the 32nd Annual Meeting of the Association for Computation Linguistics,

1994). In the 9th paragraph of the Office Action, the Examiner rejects claims 2, 7, 15 and 26

under 35 USC 103(a) as allegedly being unpatentable over Su in view of Takashi. These

rejections are respectfully traversed.

As amended, independent claims 1, 6, and 12 recite a system, method and apparatus

containing elements similar to:

> building a vocabulary comprising tokens extracted from a text corpus; and
> iteratively identifying compounds having a plurality of lengths within the text corpus,
> each compound comprising a plurality of tokens, comprising:
> > selecting n-grams having a same length that is less than the length of n-
> > grams selected during a previous iteration;

evaluating a frequency of occurrence for one or more *n*-grams having a
same length in the text corpus, each *n*-gram comprising at least one
token selected from the vocabulary;

determining a likelihood of collocation for one or more of the *n*-grams
having a the same length; and

adding a subset of *n*-grams having a high likelihood as compounds to the
vocabulary and rebuilding the vocabulary based on the added
compounds.

Independent claims 13, 24 and 36 have been amended to recite a system, method and

apparatus containing elements similar to:

iteratively specifying a limit on a number of tokens per compound for an iteration and
decreasing the limit for a subsequent iteration; and

iteratively evaluating compounds within a text corpus, comprising:

determining a number of occurrences of one or more *n*-grams within the
text corpus, each *n*-gram comprising up to a maximum number of
tokens, which are at least in part provided in a vocabulary for the
text corpus;

identifying at least one *n*-gram comprising a number of tokens equal to the
limit for the iteration based on the number of occurrences and
determining a measure of association between the tokens in the
identified *n*-gram; and

adding each identified *n*-gram with a sufficient measure of association to
the vocabulary as a compound token and rebuilding the vocabulary
based on the added compound tokens.

These features support iteratively updating a vocabulary of tokens based on the

identification of compounds. A vocabulary comprised of tokens from a text corpus is built or

provided. Compounds having a plurality of lengths are iteratively identified from a text corpus

and evaluated by selecting *n*-grams having a same length that is less than the length selected in a

previous iteration. In each iteration, a subset of *n*-grams having a high likelihood are identified as

compounds and added to the vocabulary. Alternately, n-grams may be identified as compounds

and added to the vocabulary based on a sufficient measure of association to the vocabulary as a

compound token in each iteration. The vocabulary is rebuilt based on the added compounds.

Su does not disclose these features. Su discloses a system for creating an automatic compound retrieval system. The system in Su is modeled as a two-class classification problem wherein the classifier is trained on features such as mutual information calculated from a training corpus.

Specifically, Su does not disclose iteratively updating a "vocabulary comprising tokens extracted from a text corpus" based on "iteratively identifying compounds having a plurality of lengths" or "iteratively evaluating compounds". In his analysis, the Examiner explicitly acknowledges that there is no disclosure of a vocabulary in Su but asserts that a vocabulary would be obvious to those skilled in the art. With respect to claims 1, 6 and 8, the Examiner further asserts that windowing bigrams and trigrams over the corpus can be construed as "iteratively identifying compounds having a plurality of lengths".

However, Su contains no disclosure or suggestion of iteratively updating a **vocabulary** by "adding a subset of n-grams having a high likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds". In his rejections of claims 1, 6, 8, 13, 24, and 36, the Examiner asserts that Su discloses adding the compound words having a high likelihood to a vocabulary. Su is limited to the disclosure of training a compound identifier based on a corpus. The portion of Su cited by the Examiner in his rejections (pg. 245, right column, 2nd paragraph, lines 6-8) merely discloses updating **distribution statistics** by discarding low frequency bigrams and outlier values from analysis of a tagged corpus. Distribution statistics are parameters used to train the compound classifier. These statistics are not equivalent to a vocabulary comprising tokens extracted from a text corpus. Therefore, Su fails disclose or suggest the update of a vocabulary by adding the n-grams having a highest likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds.

With respect to claims 13, 24, and 36, Su similarly fails to disclose "adding each identified *n*-gram with a sufficient measure of association to the vocabulary as a compound token" and "rebuilding the vocabulary based on the added compound tokens".

Claims 1, 6 and 8 have been amended to recite the additional limitations of cancelled claims 2 and 7 reciting an iteration "selecting n-grams having a same length that is less than a length of n-grams selected during a previous iteration", thus limiting the element "iteratively identifying compounds having a plurality of lengths". Claims 13, 24 and 36 similarly recite elements from cancelled claims 15 and 26 for "specifying the limit comprising a plurality of tokens per compound and subsequently **decreasing** the limit comprising a lesser plurality of tokens per compound". In his rejection of cancelled claims 2, 7, 15 and 26, the Examiner asserts that Takashi teaches a forward iteration (i.e. increasing the number of tokens in a n-gram) which the Examiner alleges is functionally equivalent to the above elements.

Assuming for the sake of argument that the Examiner's characterization of Takashi is correct, a forward iteration is not equivalent to the claimed invention. In his rejection of claims 2, 7, 15 and 26, the Examiner states that the iteration disclosed in Takashi would provide equivalent results as evident by the same word pairs being formed based on the size of the n-gram, the word pairs providing the same calculated likelihood. In this analysis, the Examiner neglects to consider that the results of the claimed invention are based on "adding a subset of *n*-grams having a high likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds". The results obtained from iteratively **rebuilding a vocabulary** with iterations based on increasing and decreasing *n*-gram size are not equivalent. This is because the vocabulary of tokens at each iteration is dependant on the previous iteration. Therefore, order of iteration does matter and iterations based on decreasing and increasing n-

gram size provide different results when iteratively adding n-grams to the vocabulary and rebuilding of the vocabulary based on the added tokens

Based on the above remarks, Applicants submit that independent claims 1, 6, 8, 13, 24 and 36 are patentably distinguishable over Su and Takashi, alone or in the combination suggested by the examiner. Claims 2, 7, 15 and 26 are cancelled. Claims 3, 11-12, 20-23 and 31-35 either depend directly or indirectly from claims 1, 6, 8, 13 and 24. Thus, Applicants submit that claims 3, 11-12, 20-23 and 31-35 are patentably distinguishable over Su for at least the reasons directed above to claims 1, 6, 8, 13 and 24.

In the 10th paragraph the Examiner rejects claims 4, 9, 16-18 and 27-29 as allegedly being unpatentable over Su in view of Manning (The MIT Press 1999). Claims 4, 9, 16-18 and 27-29 either directly or indirectly depend from claims 1, 6, 8, 13 and 24. Manning does not remedy the deficiencies of Su and Takashi, nor does the Examiner suggest that Manning does. Thus, Applicants submit for at least the reasons above claims 4, 9, 16-18 and 27-29 are patentably distinguishable over Su, Takashi, and Manning alone or in the combination suggested by the Examiner.

REMARKS

On the basis of the above, Applicants respectfully submit that the pending claims are patentable over the cited art. The early allowance of all claims herein is requested. If the Examiner believes that direct contact with the Applicants' attorney will advance the prosecution of this case, the Examiner is encouraged to contact the undersigned as indicated below.

Respectfully Submitted,
Franz, et al.

Date: June 19, 2007   By: /Brian Hoffman/
Brian M. Hoffman, Reg. No. 39,713
Attorney for Applicant
Fenwick & West LLP
801 California Street
Mountain View, CA 94041
Tel.: (415) 875-2484
Fax: (415) 281-1350